Brainwave: A Multidisciplinary Journal (ISSN: 2582-659X), Vol. 3, Special Issue, March 2022, pp. 79-84,

A Simple but Accurate Model for Diabetes Prediction Using Logistic Regression

Sonjoy Dhar

Student, Brainware University, email:pappusonjoy@gmail.com

Jayanta Aich

Assistant Professor, Brainware University, email:aichjayanta9@gmail.com

Abstract

Nowadays, there is little scope to found a house in India where no member in that house is affected with diabetes. This is happening not only in India, it is now a worldwide problem. This disease is responsible for various kinds of another disease including a heart attack and can cost a lot of money. But it is not possible to cure completely. Early detection of diabetes is possible which can help our doctors to treat patients well. So, we can see that early detection of diabetes is very much important. Early diagnosis and prediction of diabetes could provide the patients with an opportunity to take the appropriate preventive and treatment possibilities. To improve the risk factors of this disease, we predict diabetes for the Pima Indian dataset which is originally from the National Institute of Diabetes and Digestive and Kidney Diseases for utilizing a logistic regression model - a machine learning algorithm. Our model finds five main predictors of diabetes which are glucose, pregnancy, body mass index (BMI), diabetes pedigree function, and age, and our target variable or outcome is diabetes. Our prediction accuracy is 77.92%, which means the model fits very well in the training sample. We assure you that our model can be applied to make one of the best predictions of diabetes.

Keywords: Logistic regression, diabetes, logit transformation, machine learning, correlation, precision, accuracy.

Introduction

Diabetes is a disease that is not contagious [1]. India carries a significant part of the understanding of global diabetes [1]. India's role in diabetes review remains low [3]. Diabetes is a widespread problem in India as more than 70% of the adult population suffers from this disease [10]. So far most of the work has been done by a limited number of organizations and individuals and is limited to certain interests [3]. About 40% of the publications on diabetes in India came from only 20 organizations between 2000 and 2009 [3]. Many important aspects of diabetes in India remain neutral [3]. For the reason behind early death, diabetes is responsible [2]. It was responsible for 571,000 deaths, the 16th leading cause of global mortality in 1990 [2]. Diabetes is projected to continue to rise globally in the first quarter of the 21st century [2]. Growth will be particularly strong in India which lead the world with a 14.3% outbreak of diabetes in 1995 [2]. Recent studies of geographical and ethnographic influences have shown that people of Indian descent are extremely prone to diabetes [2]. An estimated 246 million adults were diagnosed with diabetes in 2007, 80% of whom live in developing countries, the largest number in the Indian subcontinent [1]. About 85-95% of diabetes is type 2 diabetes [1]. It is estimated that by 2025, approximately 380 million adults worldwide will have diabetes [1]. Type 2 diabetes, due to the unprecedented rate of urbanization, results in environmental and lifestyle changes [1]. The World Health Organization estimates that the urban population in developing countries will increase from 1.9 billion in 2000 to 3.9 billion in 2030 [1]. Chronic diseases, such as diabetes and cardiovascular disease (CVD), are a primary challenge for the healthcare system [1]. There have been 3068 letters from India on diabetes which is only 1.04% [4]. The major publishing houses were: All India Institute of Medical Sciences, New Delhi which has contributed the greatest number of papers (276) to the Diabetes Research Center. 126 including Chennai [4]. Other major institutions include Annamalai University, Annamalai Nagar (122), Postgraduate Institute of Medical Education and Research, Chandigarh (108), Madras University, Chennai (98), Madras Diabetes Research Foundation, Chennai (83), Jawaharlal

Nehru University, (70) and Central Pharmaceutical Research Institute, Lucknow (66) [4]. An analysis of the author's profile showed that V. Mohan and his colleagues led 174 papers from DRC, Chennai, and one after that [4]. With 173 papers from Ramachandran Madras Diabetic Research Foundation, Chennai [4]. At present, there are as many doctors as there are in India Help solve diabetes problems (including specialization in
Endocrinologist) stunned young and educated Available in endocrinology only in super specialization Level after postgraduate degree [5]. There are very few Endocrinology seats in Government Medical College and Only a few super specializations in endocrinology are offered, including the All-India Institute of Medical Sciences Delhi, Postgraduate Institute of Medical Education and Research, Chandigarh; Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, Banaras Hindu University Institute of Medical Sciences, Varanasi; And Andhra Medical College and Osmania Medical College, Hyderabad [5]. Anyone other than the medical college can achieve a Diploma of National Board from a private hospital Here, too, seats are as limited as a few hospitals Apollo and Sir Ganga Ram Hospital, New Delhi Approved by the government to provide this training [5]. Inside India, several research institutes, hospitals, and Professional firms are involved in a variety of management training and fellowship programs for physicians and paramedical staff in the prevention and treatment of diabetes [5]. Some of these are the Diabetes Education and Training M.V. with the WHO Collaborative Research Center. Hospital, Chennai; All India Institute of Diabetes and Research College, Ahmedabad, S.K. Center for Diabetes Research and Education, Calcutta; Manipal Hospital (Department of Diabetes and Endocrinology), University of Mumbai (Faculty of Medicine); All India Institute of Medical Sciences (Department of Endocrinology and Metabolism), New Delhi; Amrita Institute of Medical Sciences, Kochi; Madras Diabetes Research Foundation, Coimbatore; And the Research Foundation and Research Society for Diabetes Research on diabetes in India [5]. Of the 313 authors who participated in T1D research in India during 1996-2019, 283 published 1–5 papers each, 20 published 6–10 papers, and 10 each published 11–32 papers [6]. Research productivity of the top 15 most productive authors varies from 5 to 32 publications per author; Together they share 62.6% (321) of the publication and 65% (2962) of the citations [6]. List of 20 productive authors in T 1 D research in India [6]. Ten authors registered their publication output above the group average of 10.7, while nine authors registered their CPP and RCI above the group average of 9.2 and 14 [6]. According to the International Diabetic Federation (IDF), Low- and middle-income countries face off the biggest burden of diabetes [7]. According to the International Diabetes Federation, about 415 million people worldwide had diabetes in 2015 and that number is expected to exceed 640 million by 2040 [8]. The third type of diabetes is gestational Diabetes is actually affected by pregnancy in pregnant women [11]. Predictive models can screen people with pre-diabetes or developmental risk helps determine optimal clinical management for diabetic patients [12]. Many researchers have created various prediction models using data mining to predict and diagnose diabetes [9].

Data and Summary Statistics

We first need to know what things are needed to predict diabetes. We need blood pressure, a number of times pregnant, body mass index, age, etc. which are the main factors to predict diabetes. These variables are related to diabetes.

Parameters	**Instances**
Total Participants	768
Age	≥21
Sex	● Male : 0 ● Female : 768
Pregnancies	Numeric

Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
Body Mass Index(BMI)	Body mass index (weight in kg/(height in m))
Outcome	● Diabetic : 268 ● Non Diabetic : 500

Table 1. Details of PIMA dataset

In our model, we use the PIMA Indian dataset which is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and population near Phoenix, Arizona. In this dataset each observation belongs to an individual female patient and descriptive information on the patient's diabetes factors, along with the various medical functions such as the number of times pregnant, plasma glucose concentration 2 hours in an oral glucose tolerance test, triceps skinfold thickness in mm, body mass index measured as weight in kg / (height in m) ^ 2 (BMI), diastolic blood pressure in mm Hg, 2-Hour serum insulin in mu U / ml, age and diabetes pedigree function. Table1 represents the details of the dataset. This diabetes pedigree function provides us information about diabetes history in the patient's relatives and the genetic bond of those relatives to the patient. In this model, our target variable or outcome is diabetes, which has 2 values1 or 0. If the result is 1 then the patient has diabetes and if 0 then otherwise. We have in total 768 female patients' data. In this dataset, we have found at least one missing value index with zero in the main five factors such as BMI, blood pressure, skin thickness, insulin, glucose. So, we replace those missing values with the corresponding mean values. We described the data using a statistical function with Python 3.0. Table2 represents descriptive statistics for all variables after mean values imputation for missing values.
In this table1 shows that count which means how many data are present, the mean value of the variables, standard deviation, minimum value, quantile percentage such as 25, 50, and 75, maximum value. There are so many models are present for predicting diabetes. We are using a logistic regression model – a machine learning technique.

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	diabetes
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.681605	72.254807	26.606479	118.660163	32.450805	0.471876	33.240885	0.348958
std	3.369578	30.436016	12.115932	9.631241	93.080358	6.875374	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	20.536458	79.799479	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	79.799479	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Table2. Descriptive statistics.

Literature study

Diabetes is such a chronic degenerative and almost irreversible disease affecting and/or damaging the circulatory system resulting from impaired insulin secretion, that many researchers are devoting their research interest to diabetes for timely predicting as well as for the preventive measure. For this purpose,

the researchers have developed their own model to predict diabetes with different accuracy. Table2 represents some related work.

Table3. Related work and their accuracy

Researcher(s)	Model used	Accuracy
Iyer [13]	Naïve Bayes	79.56%
Sisodia et. al. in [14]	Decision Tree, Naïve Bayes, and SVM on PIDD	76.30%
Ram D. Joshi et. al. [12]	Logistic Regression	78.26%
Tarun [15]	PCA, REP and SVM	95.42%

Although from table3, it is observed that the accuracy of 95.42% is obtained according to Tarun et. al., but the method used a complex combination of PCA, REP, and SVM, whereas our model uses a very simple technique with the accuracy of 77.92%.

Proposed Model

In our proposed model, we have used a simple logistic regression technique for the prediction of diabetes.

Brief description of Logistic Regression Techniques:

Logistic Regression (LR) is a supervised learning classification technique, used when the dependent variable 'Y' is dichotomous or binary in nature and also discrete/categorical. But no specific rules or assumptions are taken for the predictors or the independent variables(X). LR allows to predict the probability of a particular categorical response, i.e., prediction is not in the form of an exact value, it is basically the probability. In this type of regression, there should be enough responses in every given category. On the contrary, if there are too many cells with no responses parameter estimates and standard error will likely blow up. Logistic distribution, which is S-shaped, is in probability for given input features. The estimated probability of features is between 0 and 1. This model uses a sigmoid function to find the probability of (Y=1) which is expressed as $P(Y=1|X) = \frac{1}{1+e^{-z}}$ where $z=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n$

Here, β_0 is intercept or initial value and $\beta_1, \beta_2,..., \beta_n$ are the coefficients of $X_1, X_2,\ldots, X_n$

The logistic regression model also called logit transformation or logit equation can be expressed as:

$\frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0+\beta_1X$

This is also the linear form of logistic regression.

The error function or loss function or Loss in this model is represented as

$\text{Loss} = \{-YP\ ,\ if\ Y = 1\ -(1-Y)(1-P)\ ,\ if\ Y = 0$

Our proposed algorithm uses the above mathematical expressions and a sequence of steps to build the model for the prediction of diabetes. Our model starts with the extensive literature study of diabetes mellitus. We have Jupiter notebook for implementation and Python programming language for coding. Machine learning algorithms-logistic regression classifications were implemented on the collected PIMA dataset in order to predict diabetes. The following figure fig.1 shows the steps to apply the machine learning algorithm and Fig.2 and Fig. 3 represent some effects of data preprocessing in the form of scattering plotting.

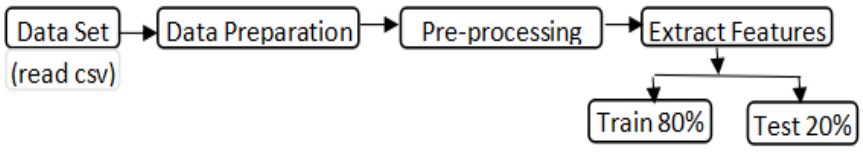
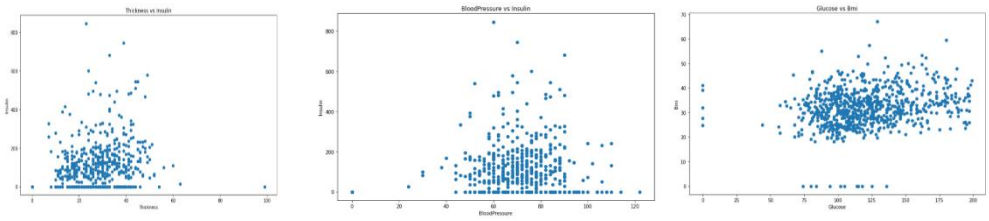


Fig1. Machine learning model



(a) (b) (c)

Fig2. Before Pre-processing (a) Thickness vs Insulin (b) Blood pressure vs Insulin (c) Glucose vs BMI

Results and Discussion

In this simple model the entire data is partitioned into two non-overlapping sets – The first set is used for training (80% and the second set is used for testing (20%) purpose. The obtained heatmap of the correlation is depicted in figure Fig.3 and discarded all the multi-collinearity for the final prediction of diabetes. The confusion matrix and the classification report are given in figure Fig 4 where the precision, recall, and f1-score are measured by the following formula:

Precision = $\frac{TP}{TP+FP}$, Recall=$\frac{TP}{TP+FN}$, f1-score=$\frac{2*Presision*Recall}{Precision+Recall}$, Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$

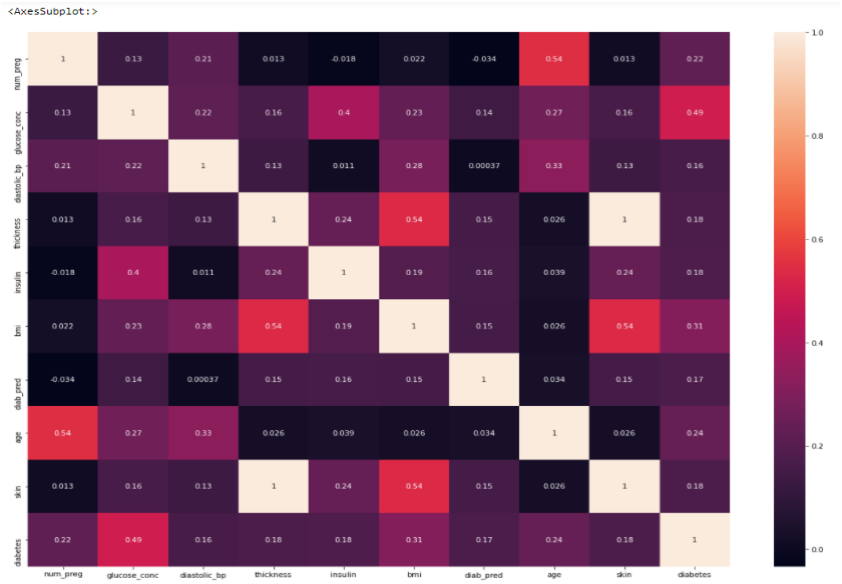
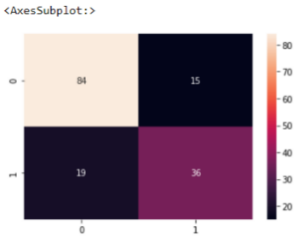


Fig3. Heatmap of the correlation of dependent and independent variables



(a)

	precision	recall	f1-score	support
0	0.82	0.85	0.83	99
1	0.71	0.65	0.68	55
accuracy			0.78	154
macro avg	0.76	0.75	0.76	154
weighted avg	0.78	0.78	0.78	154

(b)

Fig4. (a) Confusion matrix (b) Classification report

Conclusion

The experimental results show good accuracy by implementing a very simple logistic regression model and it can be used for the early prediction of diabetes. This paper will also be helpful for the researchers to implement other machine learning algorithms in a simple fashion.

References

1. Ramachandran, A., & Snehalatha, C. (2009). Current scenario of diabetes in India. *Journal of diabetes*, *1*(1), 18–28.

2. Arunachalam, S., & Gunasekaran, S. (2002). Diabetes Research in India and China Today: From Literature-based Mapping to Health-care Policy. Current Science, 82, 1086-1097.

3. Unnikrishnan, R., & Mohan, V. (2020). Whither diabetes research in India today?. Diabetes & metabolic syndrome, 14(3), 195–198.

4. Ratnakar, A., & Satyanarayana, K. (2007). Diabetes research in India--a citation profile. *The Indian journal of medical research*, *125*(3), 483–487.

5. Gupta, B., Kaur, H., & Bala, A. (2011). Mapping of Indian Diabetes Research during 1999-2008: A Scientometric Analysis of Publications Output. *DESIDOC Journal of Library & Information Technology*, 31(2)..

6. Dayal D, Gupta BM, Gupta S.(2021).Quantitative and qualitative assessment of Indian research yield in type 1 diabetes during 1996–2020. *J Diabetol*,*12*,28-35.

7. Gupta, B. M., & Bala, A. (2013). Epilepsy Research in India: A Scientometric Analysis of Publications Output during 2002-11. *Annals of neurosciences*, 20(2), 71–78.

8. Papatheodorou, K., Banach, M., Bekiari, E., Rizzo, M., & Edmonds, M. (2018). Complications of Diabetes 2017. *Journal of diabetes research*, 2018, 3086167.

9. Zhu, C., Idemudia, C.U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked.*

10. Tigga, N.P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*, 167, 706-716.

11. Md. Maniruzzaman; Md. Jahanur Rahman; Benojir Ahammed; Md. Menhazul Abedin. (2020) Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm.*Springer, 8*, 1-7.

12. Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *International journal of environmental research and public health, 18*(14), 7346

13. Iyer A, Jeyalatha S, Sumbaly R.(2015). Diagnosis of diabetes using classification mining techniques. *Int J Data Min Knowl Manag Process (IJDKP)*,*5*(1).

14. Sisodia, D., Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science 132*, 1578-1585.

15. Haldiyal, Tarun & Mishra, Pawan. (2014). Analysis and Prediction of Diabetes Mellitus Using PCA, REP, and SVM. *International Journal of Engineering and Technical Research (IJETR). 2.* 164-166.